Downloaded from orbit.dtu.dk on: May 05, 2023

DTU Library

Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain

Chabot-Leclerc, Alexandre; MacDonald, Ewen; Dau, Torsten

Published in:
Journal of the Acoustical Society of America

Link to article, DOI:
10.1121/1.4954254

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Chabot-Leclerc, A., MacDonald, E., & Dau, T. (2016). Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain. *Journal of the Acoustical Society of America*, *140*(1), 192–205. https://doi.org/10.1121/1.4954254

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

CrossMark

Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain

Alexandre Chabot-Leclerc, Ewen N. MacDonald, and Torsten Dau[a)]
Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, DK-2800, Kongens Lyngby, Denmark

(Received 25 September 2015; revised 1 June 2016; accepted 8 June 2016; published online 13 July 2016)

This study proposes a binaural extension to the multi-resolution speech-based envelope power spectrum model (mr-sEPSM) [Jørgensen, Ewert, and Dau (2013). J. Acoust. Soc. Am. **134**, 436–446]. It consists of a combination of better-ear (BE) and binaural unmasking processes, implemented as two monaural realizations of the mr-sEPSM combined with a short-term equalization-cancellation process, and uses the signal-to-noise ratio in the envelope domain (SNR_{env}) as the decision metric. The model requires only two parameters to be fitted per speech material and does not require an explicit frequency weighting. The model was validated against three data sets from the literature, which covered the following effects: the number of maskers, the masker types [speech-shaped noise (SSN), speech-modulated SSN, babble, and reversed speech], the masker(s) azimuths, reverberation on the target and masker, and the interaural time difference of the target and masker. The Pearson correlation coefficient between the simulated speech reception thresholds and the data across all experiments was 0.91. A model version that considered only BE processing performed similarly (correlation coefficient of 0.86) to the complete model, suggesting that BE processing could be considered sufficient to predict intelligibility in most realistic conditions.
© *2016 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4954254]

[FJG] Pages: 192–205

I. INTRODUCTION

Cherry (1953) coined the term "cocktail party problem" to describe the ability of listeners to "recognize what one person is saying when others are speaking at the same time." It is known that this ability is typically improved if the listeners can use both of their ears, relative to either ear alone, and if the target and maskers are spatially separated. Various models have been designed to disentangle which part of this binaural advantage can be attributed to a selection process between left versus right ear (i.e., a "better-ear" process), a "purely" binaural process where the signals from both ears interact, or a combination of both. The models typically focused on a few aspects affecting speech intelligibility at a time, such as the spatial separation of the target and the maskers, the effects of reverberation on the target or on the maskers, the role of temporal fluctuations in the masker, and the effects of multiple interferers. None of the models can account for all of these aspects at once. In the current study, recent advances in monaural intelligibility predictions are combined with binaural modeling approaches in an attempt to provide a model that can account for all of the aforementioned aspects.

Binaural intelligibility models usually combine a monaural model with some form of binaural processing to capture binaural cues attributed to head shadows and binaural interactions (Bronkhorst and Plomp, 1988). When a masker is placed elsewhere than in front of the listener, the head casts an acoustical shadow on the side opposite to the source. If the target is placed towards the ear that is in the shadow, the target-to-interferer ratio (TIR) is improved for that ear, yielding a better ear (BE), which helps the listener understand the target. These level cues are denoted as interaural level differences (ILDs). Correspondingly, different source azimuths produce different interaural time differences (ITDs). Binaural interactions rely on the ITD between target and maskers to facilitate their segregation, denoted as "binaural unmasking" (BU). The equalization-cancellation (EC) theory (Durlach, 1963) suggests that binaural unmasking can be explained by the ability of the central auditory system to "cancel" the interferers, effectively maximizing the target-to-interferer ratio.

A. Models with independent processing of ILDs and ITDs

Binaural models that predict intelligibility in spatial conditions tend to consist of a combination of two components that realize the BE and the BU processes. For example, the model of Lavandier and Culling (2010) first evaluates the BE contributions by selecting the best long-term target-to-interferer ratio for each peripheral channel, using stationary speech-shaped noise (SSN) convolved with the binaural impulse responses between the listener and the sources as the target and masker "probe signals," and combining them using the speech intelligibility index (SII) weights (ANSI, 1997). The BU path evaluates the binaural masking level difference (BMLD) in each channel using an equation based on the EC concept, which incorporates the ITDs of the target and masker, as well as the interaural coherence of the masker (Culling *et al.*, 2005). The BMLD values are also combined

[a)]Electronic mail: tda@elektro.dtu.dk

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 70.113.67.182 On: Sun, 17 Jul 2016 23:24:33

using the SII weights, and then summed with the BE to yield the overall binaural advantage, relative to the colocated condition. Their model could account for conditions with an anechoic target and a noise masker placed at different distances and azimuths in three different rooms. However, because the model considers anechoic targets only, it cannot capture the effects of reverberation on the target itself because reverberation does not strongly affect the envelopes of the convolved SSN probe signals. Furthermore, it is limited to stationary maskers and thus cannot account for intelligibility changes due to amplitude modulations in the maskers because the model only considers the long-term properties of the signals.

This model of Lavandier and Culling (2010) was expanded to include head shadow and multiple stationary maskers in anechoic (Jelfs *et al.*, 2011) and reverberant conditions (Lavandier *et al.*, 2012). Those two model versions used binaural room impulse responses (BRIRs) directly to calculate the TIRs and BMLD values. Although those extended model implementations are computationally more efficient and have more predictive power than the previous one, they still have the same inherent limitations, i.e., they cannot account for any release from masking due to modulations in the maskers and cannot describe effects of temporal smearing of the target at low direct-to-reverberant ratios. Those are similar to the limitations of the SII, on which those models are based; the models would predict good intelligibility at infinite SNRs but low direct-to-reverberant ratios, which is in contrast to the reduced intelligibility observed in such conditions.

Collin and Lavandier (2013) proposed another extension of the original work of Lavandier and Culling (2010) to account for the effects of modulated interferers, whereby the BE and BU calculations are performed in short-time frames of 12 ms on the filtered signals, rather than directly on the BRIRs. The short-time frames are averaged over the duration of the signals, similar to the processing in the extended speech intelligibility index (ESII) (Rhebergen and Versfeld, 2005). Collin and Lavandier used an SSN target, rather than speech, because it was assumed that gaps in the speech would produce negative TIRs even though they carry information that should contribute positively to the intelligibility. Collin and Lavandier (2013) varied the masker distance as well as its modulation depth using either stationary SSN, 1-, 2-, or 4-talker modulated SSN. The model was demonstrated to qualitatively account for the correct trends in the data for different masker distances and modulation depths, although measured and simulated effects were small (from less than 1 dB to about 2 dB). However, while the short-time approach seemed successful when predicting intelligibility in modulated maskers, it would fail to account for the effects of reverberation on the speech because SSN is used instead of speech for the target signal.

B. Models combining the SII and the EC concept

In the binaural speech intelligibility model (BSIM) (Beutelmann *et al.*, 2010), which is a revision and simplification of the original implementation (EC/SII; Beutelmann and Brand, 2006), the BU process is implemented as a frequency-independent equalization and cancellation (Durlach, 1963) of the long-term signals received at each ear. The SII is then used to evaluate the intelligibility based on the effective TIR in each frequency band. The reference SII value corresponding to the speech reception threshold (SRT) is selected only once for all conditions and is defined as the SII predicting 50% intelligibility for the monaural presentation of the Oldenburg Sentence Test in noise (Wagener *et al.*, 1999). The BSIM could predict SRTs of normal-hearing (NH) listeners in conditions with colocated target and stationary SSN maskers, as well as with spatially separated target and maskers, in anechoic conditions and in three different rooms (a listening room, a classroom, and a church). Beutelmann *et al.* (2010) also extended the BSIM to account for fluctuating maskers by computing the SII after EC processing in short time windows with an effective length of 12 ms. The extension was named short-term BSIM (stBSIM). The stBSIM could account for the release from masking due to modulated maskers in anechoic conditions, but was less accurate when reverberation was introduced; the mean differences between predicted and observed SRTs varied between −4.1 and −2.7 dB. Furthermore, similar to the model of Lavandier and Culling (2010) and its extensions, the stBSIM cannot account for the effects of reverberation on the speech itself because it cannot separate the useful from the detrimental part of the speech.

Rennies *et al.* (2011) proposed several modifications of the long-term BSIM to better account for the deleterious effect of reverberation: (1) an extension based on the modulation transfer function, (2) a compensation factor based on the room "definition" (a room acoustical property), and (3) a separation of the speech signal based on the useful and detrimental parts. Extension (2) provided the best fit of the three models in anechoic and reverberant conditions with stationary maskers. Although the proposed modifications increased the predictive power of the model, they also reduced its generality because the model now required access to the room impulse response in addition to the speech and noise signals.

Wan *et al.* (2010) introduced an application of the EC model of Durlach (1963), which they later denoted as the steady-state EC model (SSEC). Their approach is similar to that of the BSIM (Beutelmann *et al.*, 2010) but differs in a few important ways: the decision device based on the SII selects the best SNR from the left ear, the right ear, or from the cancelled pathway for each frequency channel, rather than from the cancelled pathway only; the EC process resolution is limited by applying frequency-independent and time-varying jitters in both amplitude and time to the output of each peripheral filter, instead of adding uncorrelated noise to each ear signal; a different SII criterion is selected for each combination of number of maskers, and type of maskers, rather than using a single SII criterion. The model was evaluated for different masker types, 1 to 3 simultaneous maskers, and different masker azimuth angles. Wan *et al.* (2010) showed that the model could predict SRTs correctly when the maskers were SSN or speech-modulated SSN, but failed when the maskers were speech or reversed speech.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 70.113.67.182 On: Sun, 17 Jul 2016 23:24:33

Wan *et al.* (2014) proposed the short-time EC model (STEC) to extend the SSEC. In contrast to the SSEC, the equalization parameters of the EC process are calculated in overlapping 20 ms windows and can vary as a function of time, which improves cancellation of the dominant masker across time. The cancelled signal is then resynthesized from the short-time windows and the SNR is calculated from the long-term spectrum. This means that only the BU process is applied in a short-time fashion and *not* the BE process. The STEC predictions were more accurate in conditions with speech-modulated SSN; however, the agreement with the data was worse than with the SSEC for reversed-speech maskers. The STEC described the spatial release from masking occurring with speech maskers slightly better than the SSEC did, but it still failed to account for the large 9 dB release from masking observed in Marrone *et al.* (2008) when two speech maskers are moved from being colocated with the target to being placed at $\pm 15^{\circ}$ azimuth angles. This may be due to differences in informational masking (IM) across the conditions. The STEC still has the same inherent limitation as the SSEC in that the model fitting has to be done for each combination of masker type and number of maskers. Further, it has never been tested in reverberant conditions.

C. Modulation-domain models

In contrast, Van Wijngaarden and Drullman (2008) extended the speech transmission index (STI) (Houtgast and Steeneken, 1973; IEC, 2003) to consider binaural hearing. The STI considers the integrity of the modulations of a reference signal (or speech) after processing as the decision metric, assessed by the modulation transfer function (MTF). The MTF can capture the effects of reverberation on speech because of the reduction in modulation in the reference signal. The binaural interaction of the binaural STI is based on interaural cross-correlograms. Van Wijngaarden and Drullman (2008) showed that the binaural STI extension could account for consonant-vowel-consonant (CVC) word scores for stationary maskers presented in multiple rooms (anechoic, a listening room, a classroom, and a large church). However, this approach is limited because it cannot be extended to more realistic conditions where the maskers are also modulated, since modulations are then coming from both the target and maskers and they can no longer be distinguished.

In order to account for different amounts of target and masker modulations, Jørgensen and Dau (2011) proposed the monaural speech-based envelope power spectrum model (sEPSM), which considers the signal-to-noise envelope power ratio (SNR_{env}) at the output of a modulation filterbank (Ewert and Dau, 2000) as the decision metric. In addition to conditions with additive maskers, the sEPSM can also account for the effects of reverberation, as well as noise reduction via spectral subtraction because it captures the increase in the masker's modulation power *after* processing. The sEPSM was extended to account for conditions with fluctuating maskers by using a "multi-resolution" process (Jørgensen *et al.*, 2013). In the corresponding multi-resolution model, the mr-sEPSM, the SNR_{env} is calculated in windows of different length [akin to the ESII of Rhebergen and Versfeld (2005)] according to the center frequency of the modulation filters. The mr-sEPSM was validated using various fluctuating noises, including cafe noise, two-band speech modulated noise, the international speech test signal (Holube *et al.*, 2010), and a reversed talker. In contrast to the SII and STI metrics, the SNR_{env} metric can account for both the effects of reverberation on the target and the masker as well as for the release from masking due to fluctuations in the maskers. However, the model has not yet been applied to spatial conditions using two-ear processing. Therefore, using the mr-sEPSM framework in a binaural model could yield a model that can account for all the aforementioned aspects of binaural speech intelligibility: the spatial separation of the target and the maskers, the effects of reverberation on the target and on the maskers, the role of temporal fluctuations in the masker, and the effects of multiple interferers.

None of the models previously mentioned can account for the deleterious effects of colocated concurrent speakers on speech intelligibility. The difference between the measured intelligibility and intelligibility predicted using energy-based model is often labeled as "informational masking."

D. Proposed modeling framework

Here, a model is proposed that combines concepts from different modeling approaches. Specifically, it integrates a short-time equalization-cancellation process (Wan *et al.*, 2014), a temporal modulation filterbank (Dau *et al.*, 1997; Ewert and Dau, 2000), the SNR_{env} metric (Jørgensen and Dau, 2011), and a better-ear process in the envelope power domain. The model was evaluated using a set of critical experimental conditions from the literature to tease apart the contributions of the decision metric, the short-time processing, the better-ear process, and the binaural unmasking for predicting intelligibility in spatial conditions. Experiment 1 focused on conditions with multiple maskers in anechoic conditions, experiment 2 considered conditions with only a single masker, but in a reverberant environment, and experiment 3 investigated a single-masker condition where only ITD but no ILD information was provided.

II. MODEL DESCRIPTION

A. Overall model structure

Figure 1 shows a sketch of the model proposed in the present study, which is an extension of the monaural mr-sEPSM (Jørgensen *et al.*, 2013). The model consists of realizations of the monaural mr-sEPSM for the left and right ear, and a "central" pathway where binaural unmasking takes place using an EC process (Wan *et al.*, 2014). In contrast to the original mr-sEPSM, the model employs a binaural processing stage. Binaural processing is limited by peripheral transduction, which does not preserve fine-structure information at high frequencies (Bernstein and Trahiotis, 1996). Peripheral transduction is therefore modeled using half-wave rectification and low-pass filtering. A binaural selection stage combines the outputs of the left, right and central pathways. The subsequent output is then converted to intelligibility using an ideal

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 70.113.67.182 On: Sun, 17 Jul 2016 23:24:33

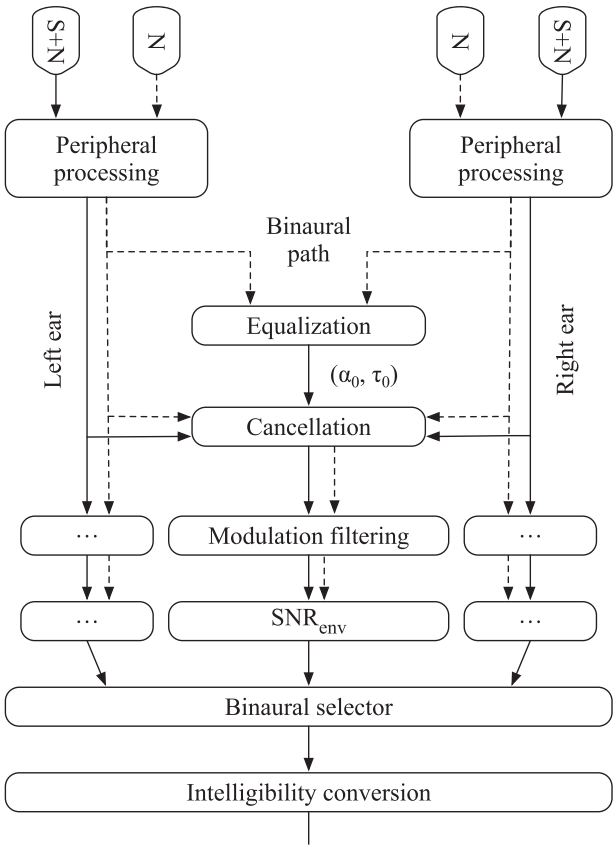

FIG. 1. Diagram of the model structure. Solid lines denote the path of the speech-plus-noise ($S+N$) mixture and the dash lines show the path of the noise alone (N). The values α_0 and τ_0 represent the optimal parameters selected by the equalization process.

observer concept. The extension to the mr-sEPSM is presented below; further details and justifications about the mr-sEPSM approach can be found in Jørgensen *et al.* (2013).

B. Monaural processing stage

The inputs of the model are the noisy speech and the noise alone for each ear. The first stage of each monaural model consists of 22 gammatone filters covering the frequency range from 63 Hz to 8 kHz with a third-octave spacing. The channels are processed further only if the level of the noisy speech for that channel is above the diffuse-field threshold in quiet (ISO, 2005). The envelope of each channel output is extracted using half-wave rectification and low-pass filtered using a fifth-order Butterworth filter with a cut-off frequency of 770 Hz (Breebaart *et al.*, 2001). Jitter in amplitude and time is applied to each envelope independently to limit the efficacy of the EC process; all jitters are zero-mean Gaussian processes with standard deviations of $\sigma_\delta = 105\,\mu$s for the time jitter and of $\sigma_\epsilon = 0.25$ (dimensionless) for the amplitude jitter (Durlach, 1963). The resulting envelopes are further processed by a modulation filterbank consisting of eight second-order Butterworth band-pass filters. A third-order low-pass filter with a 1 Hz cut-off frequency is applied in parallel, which completes the filterbank. Conceptually, this filter can be considered as the lowest frequency band in the filterbank. Only modulation filters with center frequencies below one-fourth of their respective peripheral-filter center frequency are used (Verhey *et al.*, 1999).

The output of each modulation filter, n, is segmented in non-overlapping rectangular windows of durations inversely proportional to the center-frequency of the respective modulation filter, e.g., the windows at the output of the 8 Hz modulation filter are 125 ms long. The power, $P_{\text{env},i}(p, n)$, of each segment, i, is defined as the variance of the segment,

$$P_{\text{env},i}(p,n) = \frac{1}{\left[\bar{E}(p,t)\right]^2/2}\overline{\left[e_i(p,n,t) - \bar{e}_i(p,n)\right]^2}, \quad (1)$$

where p is the corresponding peripheral filter, $E(p, t)$ is the envelope at the output of the peripheral filter, $e_i(p, n, t)$ is the envelope at the output of the modulation filter for the segment i, t is time, and the overbar indicates the average over time. $\bar{e}_i$ is the average over a time segment, i, of varying duration according to the center frequency of the modulation filter. $\bar{E}$ is averaged over the whole sentence duration. The lower limit of the envelope power is set to -30 dB relative to 100% amplitude modulation.

The $\text{SNR}_{\text{env},i}$ for each segment is computed from the envelope power of the noisy speech and the noise alone,

$$\text{SNR}_{\text{env},i}(p,n) = \frac{P_{\text{env},S+N,i}(p,n) - P_{\text{env},N,i}(p,n)}{P_{\text{env},N,i}(p,n)}, \quad (2)$$

where $S+N$ denotes the noisy speech and N denotes the noise alone.

C. Binaural processing stage

The binaural unmasking stage is implemented as described in Wan *et al.* (2014). The jittered peripheral envelopes from the monaural stages are used as inputs to the EC process. The EC processing is assumed to be independent in each channel, and performed in short overlapping time frames. A time-frequency unit is denoted as $U(p, k)$, where p again denotes the peripheral filter, and k is the kth frame, which differs from the ith segment of the modulation-domain multi-resolution process. Each frame, k, is 20 ms, whereas the multi-resolution segments, i, can vary in duration. The overlap between frames is 50% (10 ms). The equalization process in each unit selects the optimal ITD, τ_0, and the optimal ILD, α_0, using the following equations:

$$\tau_0(p,k) = \arg\max_{\tau}\{\rho_{p,k}\}, \quad |\tau| < \frac{\pi}{\omega_p},$$
$$\alpha_0(p,k) = \sqrt{\frac{F_{N,L}(p,k)}{F_{N,R}(p,k)}}, \quad (3)$$

where $\rho_{p,k}$ is the normalized cross-correlation function of the left and right ears within the unit, $F_{N,L}(p, k)$ and $F_{N,R}(p, k)$ are the masker energy for the left and right ear, respectively, and ω is the center frequency of channel p. The unmasked output, $Y_{p,k}(t)$, for the unit $U(p, k)$ after cancellation is calculated as

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 70.113.67.182 On: Sun, 17 Jul 2016 23:24:33

$$Y_{p,k}(t) = W_k(t)\left\{ \frac{1}{\sqrt{\alpha_0(p,k)}} E_L(p,t)\left(t + \frac{\tau_0(p,k)}{2}\right) - \sqrt{\alpha_0(p,k)} E_R(p,t)\left(t - \frac{\tau_0(p,k)}{2}\right)\right\}, \quad (4)$$

where the subscripts L and R denote the left and right ear, respectively, and $W_k(t)$ is a rectangular window function for the frame k, which can be expressed as

$$W_k(t) = \begin{cases} 1, & (k*10)\,\text{ms} \le \text{t} \le (k*10) + 20\,\text{ms}, \\ 0, & \text{otherwise}. \end{cases} \quad (5)$$

Subsequently, the binaural signal, B_p, is reconstructed for each channel by summing over all overlapping frames

$$B_p(t) = \frac{1}{2}\sum_k Y_{p,k}(t). \quad (6)$$

The unmasked outputs for the noisy speech, $B_{S+N,p}$, and the noise alone, $B_{N,p}$, are then used as inputs to the modulation filtering stage of the mr-sEPSM, and, subsequently, to the SNR_{env} calculation. This yields $\text{BU-SNR}_{\text{env}}(p, n, t)$, a binaurally unmasked SNR_{env}, for each peripheral channel, modulation channel, and multi-resolution frame.

D. Binaural selection stage

The binaural selection device selects the best SNR_{env}, denoted as the "binaural SNR_{env}" ($\text{B-SNR}_{\text{env}}$), between the better-ear SNR_{env}, ($\text{BE-SNR}_{\text{env},i}$) and the binaurally unmasked $\text{SNR}_{\text{env},i}$ ($\text{BU-SNR}_{\text{env},i}$) for each multi-resolution segment [note that the (p, n, t) indexing has been omitted for the sake of brevity],

$$\text{B-SNR}_{\text{env},i} = \max(\text{BE-SNR}_{\text{env},i}, \text{BU-SNR}_{\text{env},i}), \quad (7)$$

where $\text{BE-SNR}_{\text{env},i}$ is the maximum between the left and right $\text{SNR}_{\text{env},i}$ for each segment

$$\text{BE-SNR}_{\text{env},i} = \max(\text{SNR}_{\text{env},L,i}, \text{SNR}_{\text{env},R,i}). \quad (8)$$

The $\text{B-SNR}_{\text{env}}$ is then averaged over all segments, I_n, of each modulation channel

$$\text{B-SNR}_{\text{env}}(p,n) = \frac{1}{I_n}\sum_{i=1}^{I_n} \text{B-SNR}_{\text{env},i}(p,n), \quad (9)$$

yielding a 9×22 array of values. The time-averaged $\text{B-SNR}_{\text{env}}$ is first combined across modulation filters

$$\text{B-SNR}_{\text{env}}(p) = \left[\sum_{n=1}^{9} \text{B-SNR}_{\text{env}}^2(p,n)\right]^{1/2} \quad (10)$$

and then across peripheral filters

$$\text{B-SNR}_{\text{env}} = \left[\sum_{p=1}^{22} \text{B-SNR}_{\text{env}}^2(p)\right]^{1/2}. \quad (11)$$

E. Decision device

The overall $\text{B-SNR}_{\text{env}}$ is converted to a sensitivity index, d', of an "ideal observer" (Jørgensen and Dau, 2011), using the relation

$$d' = k(\text{B-SNR}_{\text{env}})^q, \quad (12)$$

where k and q are parameters independent of the experimental conditions. d' is converted to intelligibility using an m-alternative forced choice decision model, combined with an unequal variance Gaussian model expressed as

$$P_{\text{correct}}(d') = \Phi\left(\frac{d' - \mu_N}{\sqrt{\sigma_S^2 + \sigma_N^2}}\right), \quad (13)$$

where Φ denotes the cumulative normal distribution. The values of σ_N and μ_N are determined by the number of response alternative, m [see the Appendix of Jørgensen and Dau (2011) for details]. For open-set paradigms, m is set to 8000, which reflects the number of words in a normal listener's vocabulary. The value of σ_S is a free parameter fixed by fitting model predictions to speech intelligibility data in a condition with a SSN masker. The percentage correct at the output of the model is denoted as B-sEPSM.

Replacing the $\text{B-SNR}_{\text{env}}$ by either $\text{BE-SNR}_{\text{env}}$ or $\text{BU-SNR}_{\text{env}}$ in Eqs. (9)–(13) provides two alternative model outputs, BE-sEPSM and BU-sEPSM, where only the better-ear or only binaural-unmasking stages are used, respectively.

III. METHODS

A. Experiment 1: Multiple maskers in an anechoic condition

1. Rationale

This experiment investigated the effects of multiple spatially distributed maskers in an anechoic condition on spatial release from masking (SRM) using the data of Hawley *et al.* (2004). They systematically measured SRTs as a function of masker azimuth, masker type, and number of maskers using the Harvard IEEE corpus (Rothauser *et al.*, 1969). The interferers were either speech (not considered in the current study because of potential differences in informational masking compared to the other conditions), reversed speech (from the same corpus), SSN, or speech-modulated speech-shaped noise (SMSSN). All maskers were matched to the spectrum of the target sentences and either one, two, or three maskers were presented at once. Hawley *et al.* showed that SRM was larger when multiple voiced interferers were located at different locations from the target, compared to conditions when a single voiced masker was presented. This suggested that a short-term binaural process is critical. Wan *et al.* (2010) and Wan *et al.* (2014) used these same data to validate their long-term and short-term EC models.

2. Experimental conditions

The Loizou (2007) recording of the Harvard IEEE corpus, sampled at 25 kHz, was used for the target material. The

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 70.113.67.182 On: Sun, 17 Jul 2016 23:24:33

SSN was also taken from Loizou (2007) and was created by filtering stationary noise to have the same long-term spectrum as the speech material. The SMSSN was created by applying the broadband envelope of a sentence from the speech corpus to the SSN. The envelope was extracted by low-pass filtering the half-wave rectified speech signal with a first-order Butterworth filter with a 40-Hz cutoff frequency (Hawley *et al.*, 2004). The stimuli were spatialized using the head-related transfer functions (HRTFs) of the HMS II artificial head (HEAD acoustics GmbH, Germany) from the AUDIS database (Blauert *et al.*, 1998), at angles of 0°, 30°, 60° or 90°. One to three maskers were located in the front (0°, 0°, 0°), to the side (90°, 90°, 90°), distributed on the right (30°, 60°, 90°), or distributed to the left and the right (−30°, 60°, 90°) [see Table I in Hawley *et al.* (2004) for the full layout]. The speech level was fixed at 65 dB sound pressure level (SPL) and the masker levels were independently set to desired SNRs, before HRTF filtering; adding maskers increased the total interferer level.

3. Simulations

Simulations obtained with the proposed model (B-sEPSM) were carried out for SNRs ranging from −24 to 12 dB in 3 dB increments. The simulated SRTs corresponded to the SNR at which the simulated intelligibility was 50%, using linear interpolation where necessary. The final SRT represented the average SRT for 30 randomly selected sets of target and maskers. The condition with a single SSN masker, colocated with the target, and spatialized using the AUDIS HRTFs was considered as the reference condition. Because word score data were unavailable in this condition, a Gaussian psychometric function, $c(x)$, with an anechoic SRT, μ_a, and a standard deviation, σ, was first fitted based on the anechoic word score data of Bernstein and Grant (2009) using

$$c(x) = \mathrm{erfc}\left(\frac{-(x-\mu_a)}{\sqrt{2}\sigma}\right)\Big/2, \tag{14}$$

where x represents the SNRs, c is the proportion correct, and erfc is the complementary Gauss error function. Then, μ_a was replaced by the SRT measured by Hawley *et al.* (2004) in the colocated condition with a single SSN masker (−3.40 dB SNR), yielding a modified psychometric function, $c'(x)$. The parameters of the model's ideal observer, k and q, were adjusted to minimize the root-mean-square error (RMSE) between the simulations obtained with the "left ear" model and the psychometric function $c'(x)$. The constants σ_s and m of the observer were fixed to 0.6 and 8000, respectively. The observer parameters were kept constant throughout experiment 1. Table I shows the summary of the ideal observer parameters and constants for all three experiments.

TABLE I. Calibrated values of the parameters k and q, and of the constants σ_s and m of the ideal observer for the different experiments.

Condition	k	q	σ_s	m
Exp. 1: Hawley *et al.* (2004)	0.82	0.31	0.6	8000
Exp. 2: Beutelmann *et al.* (2010)	0.04	1.42	0.9	50
Exp. 3: Lőcsei *et al.* (2015)	1.14	0.235	0.6	8000

B. Experiment 2: Single masker in reverberant conditions

1. Rationale

In contrast to experiment 1, experiment 2 considered the effects of a single masker of different types, but in reverberant conditions, using the data of Beutelmann *et al.* (2010). They measured binaural SRTs in a combination of four different rooms, three target-masker azimuth separations, and three masker types. The speech material was the Oldenburg Sentence Test in noise (Wagener *et al.*, 1999), which consists of a closed set of meaningful five-word sentences. The rooms included an anechoic room, a standard IEC listening room (not used in the current study), a typical classroom ($7 \times 6.9 \times 3.2$ m^3, approximately 210 m^3) and a church (outer dimensions: $63 \times 32 \times 22$ m, approximately 22 000 m^3). The interferers were either stationary SSN (stationary), 20-talker babble (babble), or single-talker modulated noise (single-talker). Beutelmann *et al.* (2010) found an effect of azimuth on the SRM (a 105° separation yielded a larger SRM than a 45° separation) and this effect was largest in the anechoic condition. They also showed that the masker type had a significant effect on the SRM and that speech intelligibility was positively correlated with the modulation depth of the masker. SRM was larger in the anechoic conditions, than in the reverberant conditions. The masker types used by Beutelmann *et al.* (2010) were similar to the ones used by Hawley *et al.* (2004) but the different room types are critical to validate that the proposed model can capture the effects of reverberation on SRM.

2. Conditions

The SSN was the one provided with the Oldenburg Sentence Test, which was created by a random superposition of the material's sentences. The multi-talker babble was the "CD101RW2" noise from the Auditec CD, created as the mixture of 20 talkers reading different texts (Auditec, 2006). The single-talker modulated noise was the "ICRA5" noise (Dreschler *et al.*, 2001). All stimuli were sampled at 44.1 kHz. The noise level was fixed at 65 dB SPL and the target level was adjusted to the desired SNRs. Both the target and masker levels were adjusted *after* HRTF filtering. The stimuli were spatialized using virtual impulse responses created with the ODEON software version 8.0 (Kongens Lyngby, Denmark; Christensen, 2005). The anechoic, the classroom and the church conditions were used. Three spatial setups were used: (1) the target and the masker were colocated 3 m in front of the listener, (2) the target was 3 m in front of the listener and the masker was 2 m away, at 105° azimuth, and (3) the target was 6 m in front of the listener and the masker was 4 m away, at −45° azimuth. In the third condition, the listener was placed close to a wall on the right side. See Beutelmann *et al.* (2010) for complete details about the listening test setup.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 70.113.67.182 On: Sun, 17 Jul 2016 23:24:33

3. Simulations

Simulations were obtained for SNRs ranging from −36 to 6 dB in 3 dB steps. The final simulated SRT was the average SRT for 30 randomly selected target and masker pairs. The reference psychometric function, p, was created following Wagener and Brand (2005),

$$p(L, \mathrm{SRT}, s) = 100 * \frac{1}{1 + \mathrm{e}^{4s(\mathrm{SRT}-L)}}, \quad (15)$$

where L represents the given SNRs, s is the slope around the 50% point, and SRT is the SNR at the 50% points. s was set to 0.18/dB according to Wagener and Brand (2005, their Table IV) for the OLSA material with an SSN masker, and SRT was the median SRT in the spatialized condition measured by Beutelmann *et al.* (2010) (−7.23 dB, their Fig. 6) for the same material. The ideal observer parameters in the proposed model were fit such as to minimize the RMSE between the "left-ear" of the model and that psychometric function. The parameters were kept constant for all other conditions in this experiment. The observer's constants, σ_s and m, were fixed to 0.9 and 50, respectively, to account for the increased redundancy in the speech material.

C. Experiment 3: ITD-only condition

1. Rationale

Experiment 3 investigated the role of the EC process using a condition where the target and masker were lateralized to the left or to the right, using a fixed interaural delay (ITD) of 687.5 μs (Lőcsei *et al.*, 2015). The speech was played in the presence of SSN that was either lateralized to the same side as the speech, denoted as condition S11, or to the opposite side, denoted as condition S01. Lőcsei *et al.* (2015) found a masking release of about 4 dB when the masker was lateralized to the opposite side. In this condition, no better-ear benefit can be expected because the signal at both ears is the same, except for a short delay used for the lateralization. Therefore, the only cues available should be interaural differences, which should be captured by the EC process.

This experiment was akin to the $S_\pi N_0$ condition often used as an example of pure-tone BMLD [see Levitt and Rabiner (1967), and Culling *et al.* (2004)]. In such a condition, listeners showed a masking release as large as 12 dB when the target tone was presented out of phase (π), compared to the in-phase presentation of the target (0; Levitt and Rabiner, 1967). Release from masking due to ITD or out-of-phase presentation has successfully been modeled for pure-tone signals (Levitt and Rabiner, 1967) and for speech signals using an EC-like process (Culling *et al.*, 2004).

2. Conditions

The speech material was the DAT corpus (Nielsen *et al.*, 2014), sampled at 48 kHz and recorded with female speakers. The DAT corpus consists of unique meaningful Danish sentences constructed as a fixed carrier sentence with two interchangeable target words. The masker was stationary noise shaped to have the same long-term spectrum as the speech material. The target level was fixed at 65 dB SPL and the masker level was adjusted to the desired SNR.

3. Simulations

Simulations were obtained for 30 randomly selected sentences and SSN maskers, and for SNRs from −12 to 9 dB in 3 dB steps. The signals were lateralized to the left or right using a fixed 33 sample delay (687.5 μs). The final simulated SRT was the average across target sentences. The ideal observer's parameters were fit to minimize the RMSE between the "left-ear" of the model and the word-scores as a function of SNR in the colocated, S11, condition, as measured by Lőcsei *et al.* (2015). The ideal observer's σ_s and m were set to 0.6 and 8000, respectively (Jørgensen *et al.*, 2013).

IV. RESULTS

A. Experiment 1: Multiple maskers in an anechoic condition

Figure 2 shows the simulated SRTs obtained with the proposed model (B-sEPSM; black squares), those obtained with the better-ear only version of the model (BE-sEPSM; dotted line) as well as the binaural-unmasking version (BU-sEPSM; dashed line) as a function of the masker(s) angle(s). Furthermore, the STEC predictions from Wan *et al.* (2014) (grey triangles) and the measured data from Hawley *et al.* (2004) (open squares) are shown. The three columns correspond to one (left), two (middle), or three maskers (right), respectively. The upper panels show data and simulations for the stationary SSN maskers, the middle panels for SMSSN maskers, and the bottom panels for reversed speech. Figure 3 is a replot of the data and predictions of Fig. 2 where the thresholds are represented in terms of a SRM relative to the condition where the target and the maskers were colocated.

Overall, there was a good agreement between the B-sEPSM simulations and the data. The Pearson correlation coefficient across all conditions was 0.91 and the prediction RMSE was 3.0 dB. For the STEC, the correlation coefficient was 0.97 the RMSE was 1.3 dB SNR. Thus, the RMSE was larger for the B-sEPSM than for the STEC but, unlike the STEC, the B-sEPSM was fit only once for all conditions. In contrast, the STEC was fit to the 90° condition for each combination of n maskers and masker type, i.e., for each subfigure of Fig. 2 (Wan *et al.*, 2014).

In the SSN condition (upper panels), the B-sEPSM simulations were slightly lower than in the data but the amount of SRM was well described for all numbers of maskers. In the SMSSN masker condition (middle panels), the B-sEPSM correctly accounted for the masker-type dependency of the SRTs in the case of the single masker. The B-sEPSM predicted an increase in SRTs with increasing number of maskers, consistent with the measured data; however, the SRTs were on average 4.76 dB larger than in the data in the condition with three SMSSN maskers. The simulated SRM was found to be the same as in the data with two SMSSN

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 70.113.67.182 On: Sun, 17 Jul 2016 23:24:33

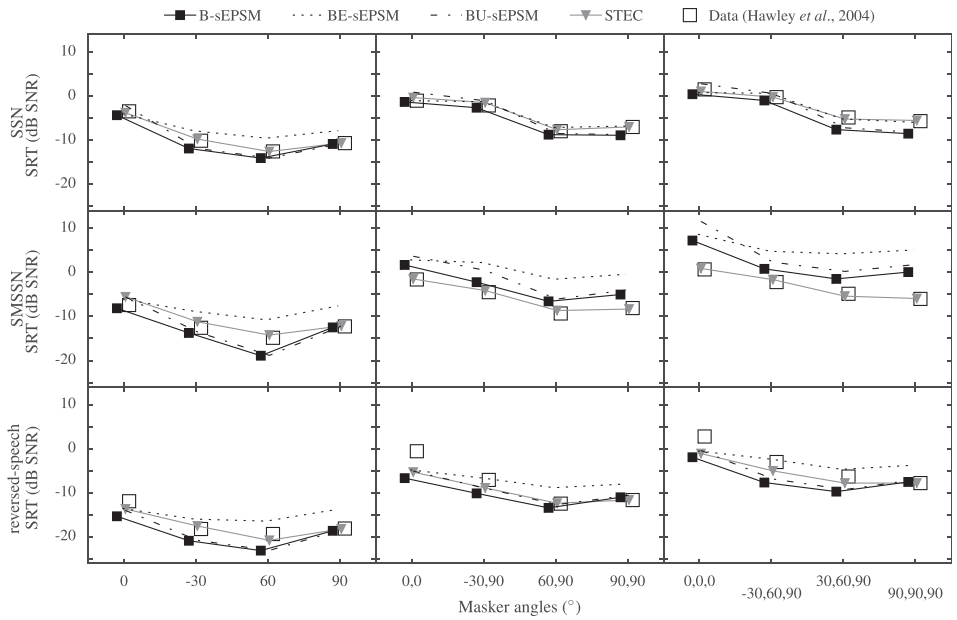

FIG. 2. Mean speech reception threshold data (open squares; Hawley *et al.*, 2004) and predictions obtained with the proposed model (black squares; B-sEPSM) and its alternate outputs, BE-sEPSM (dotted lines) and BU-sEPSM (dashed lines) as a function of masker(s) angle(s) for SSN masker(s) (upper panels), SMSSN masker(s) (middle panels), and reversed speech (bottom panels). For comparison, STEC model predictions are shown as grey triangles for reference (Wan *et al.*, 2014). The left panels show the condition with one masker only; the middle panels show the conditions with two maskers; and the right panels show the conditions with three maskers.

maskers, but was larger by about 4 dB with three maskers when all maskers were at different locations. The B-sEPSM predicted SRTs up to 8 dB higher in the three SMSSN maskers condition compared to the three SSN maskers condition. This is in contrast to the data, where the SRTs differed, on average, by only 1 dB between the SSN and SMSSN conditions when there were multiple maskers. Thus, the addition of a second or third SMSSN masker decreased the SNR_{env} more than the addition of SSN maskers. In the reversed-speech masker condition (lower panels), the B-sEPSM simulated SRTs were below the measure ones. However, as in the data, the simulated SRTs increased with the number of maskers, suggesting that the B-sEPSM could correctly account for intelligibility as a function of the number of reversed-speech maskers.

B. Experiment 2: Single masker in reverberant conditions

Figure 4 shows the measured SRTs from Beutelmann *et al.* (2010) (open squares), together with the B-sEPSM predictions (black squares), the simulations obtained with the better-ear (BE-sEPSM; dotted lines) and binaural-unmasking (BU-sEPSM; dashed lines) versions of the model as a function of the masker azimuth. Furthermore, the stBSIM predictions (grey bullets; replotted from Beutelmann *et al.*, 2010) are shown for comparison. The three columns correspond to the anechoic, classroom, and church conditions, respectively. The upper panels show data and predictions for the stationary masker, the middle panels show the corresponding results for the babble masker, and the bottom panels show the results obtained for the single-talker modulated noise masker.

Overall, there was a good agreement between the predictions and the data. The B-sEPSM Pearson correlation coefficient across all conditions was 0.91 and the average prediction RMSE for the B-sEPSM was 6.5 dB. In contrast, the Pearson coefficient for the stBSIM was 0.89 and the RMSE was 3.65 dB.

In the anechoic condition (left panels), the B-sEPSM produced a larger SRM than that found in the data when the masker was stationary noise or single-talker noise. A similar SRM as in the data was found when the maskers were babble noise. In the classroom condition (middle column), the B-sEPSM accurately accounted for the SRM but there was a negative offset for all masker types. In the church condition

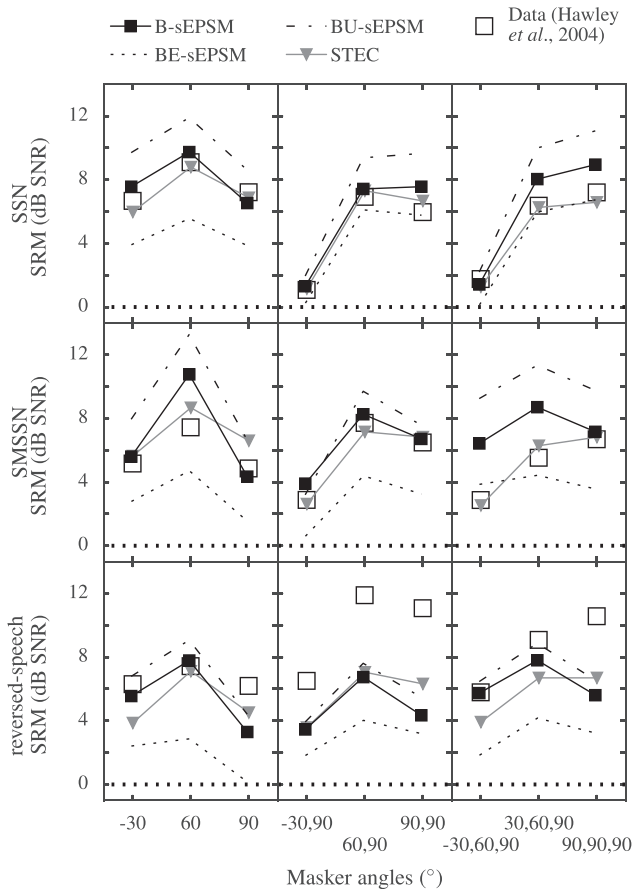

FIG. 3. Replot of the data and predictions of Fig. 2 as SRM relative to the colocated condition.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 70.113.67.182 On: Sun, 17 Jul 2016 23:24:33

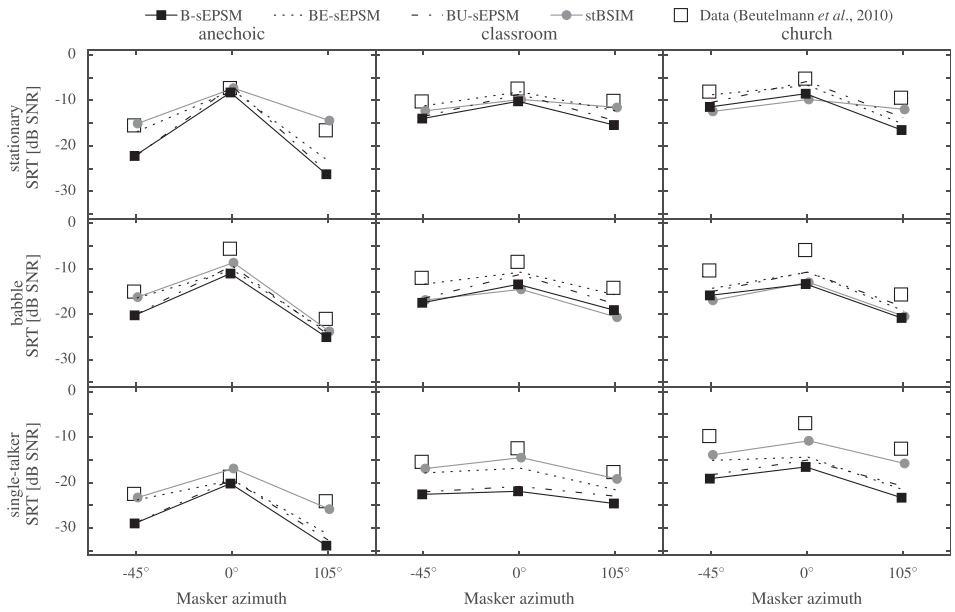

FIG. 4. Median speech reception thresholds data measured by Beutelmann *et al.* (2010) (open squares), B-sEPSM predictions (black squares), BE-sEPSM predictions (dotted lines), BU-sEPSM predictions (dashed lines), and stB- SIM predictions (grey bullets; Beutelmann *et al.*, 2010) as a function of the azimuth of stationary SSN (upper panels), babble noise (middle panels), or a single-talker modulated noise (bottom panels).

(right column), the SRM was also correctly accounted for by the B-sEPSM, except for a negative offset which was largest for the single-talker babble noise. Overall, the B-sEPSM offset seemed to be partly due to the BU-sEPSM contributions, which were consistently lower than the BE-sEPSM contributions to the overall SNR_{env}. Nonetheless, the large offset observed in the reverberant conditions for all maskers was due to the particulars of the intelligibility transformation for the B-sEPSM. The sEPSM framework is sensitive to the type of SSN used in the reference condition; white-noise filtered to have the same long-term average spectrum as speech, and SSN created by the random superposition of speech signals yield different amounts of masking release. A smaller offset could be obtained if the ideal observer was fit to the B-sEPSM, rather than to the "left-ear" SNR_{env}, but the resulting binaural model could not be analyzed in terms of the benefit compared to one ear alone in the colocated condition.

Figure 5 is a replot of the data and predictions from Fig. 4 as spatial release from masking relative to the colocated condition. The data from Beutelmann *et al.* (2010) showed that SRM decreased with increasing amounts of reverberation, probably as the result of reduced head shadow effect which decreases the BE benefit (Lavandier and Culling, 2010; Plomp, 1976). Reverberation also decorrelates the signals that reaches both ears, which reduces the efficacy of the EC process (Lavandier and Culling, 2007). These effects were captured by the BE-sEPSM and the BU-sEPSM outputs, respectively, and therefore, by the B-sEPSM, for all masker types, as shown by the correctly predicted SRM (cf. Fig. 5, second and third columns).

Release from masking with a fluctuating masker, relative to a stationary masker, was also reduced in the presence of reverberation; the SRT in the colocated single-talker masker church condition was about 12 dB higher (−18.7 to −6.95 dB SNR) than in the anechoic condition. None of the models accurately predicted this large SRT increase; the B-sEPSM predicted an increase of 3.65 dB and the stBSIM an increase of 6.10 dB.

C. Experiment 3: ITD-only condition

The left panel of Fig. 6 shows the measured SRTs (open squares) from Lőcsei *et al.* (2015), the B-sEPSM predictions (black squares), as well as the predictions from the better-ear-only version of the model (BE-sEPSM; dotted line and diamonds) and the binaural-unmasking version (BU-sEPSM; dashed line and circles). Target and masker were colocated to the left in the S11 condition. In the S01 condition, the

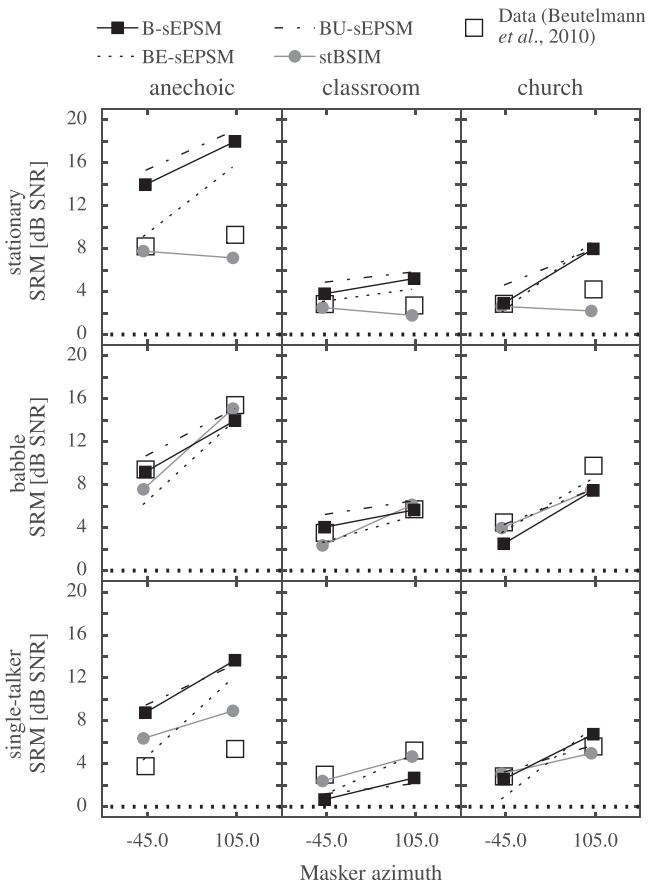

FIG. 5. Replot of the data and predictions of Fig. 4 as spatial release from masking relative to the colocated condition.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 70.113.67.182 On: Sun, 17 Jul 2016 23:24:33

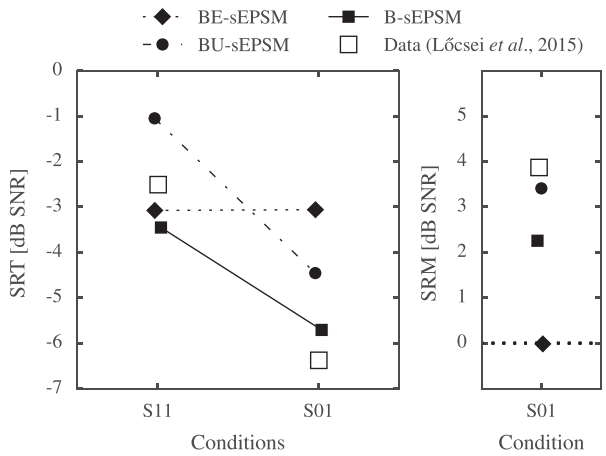

FIG. 6. The left panel shows speech reception threshold data (open squares) from Lőcsei *et al.* (2015), B-sEPSM predictions (black squares), BE-sEPSM predictions (solid lines), and BU-sEPSM predictions (dashed lines) as a function of condition. In condition S11, both target and maskers are lateralized to the left and in S01, the target is lateralized to the left and the masker to the right. The right panel shows the same data and predicted, replotted as spatial release from masking relative to the S11 condition.

target was lateralized to the left and the masker was lateralized to the right. The right panel shows the same data and predictions replotted as SRM relative to the S11 condition. The B-sEPSM predicted SRTs lower than the measured ones in the S11 condition even though the model was fitted to that condition. This can be explained by the fact that the B-sEPSM was fit to the "left-ear" output only, rather than to the complete model output. Consequently, there seems to be a small advantage inherent to the binaural presentation in this condition, compared to the monaural presentation. The B-sEPSM produced an SRM of about 2 dB, compared to the 4 dB observed in the data. The BE-sEPSM output predicted no release from masking because there was no BE benefit possible; the masker was identical in both ears except for the fixed delay which is about an order of magnitude shorter than all processing windows in the model. In contrast, the BU-sEPSM output could account for all the SRM observed in the data.

V. DISCUSSION

This study described a binaural extension of the mr-sEPSM model framework, which combined monaural implementations of the mr-sEPSM with the EC model implementation of Wan *et al.* (2014). The regular mr-sEPSM process was applied to the envelopes at the output of the EC process, and a selection stage selected the best SNR_{env} from the left-ear, the right-ear—equivalent to better-ear processing—or the binaural unmasking pathway. The model was validated against the data of Hawley *et al.* (2004), Beutelmann *et al.* (2010), and of Lőcsei *et al.* (2015). Overall, the correlation coefficients between simulated and measured SRTs were equal to 0.91. See Table II for a summary of all model performances.

A. Comparison to other modeling approaches

Both the proposed model and the STEC (Wan *et al.*, 2014) had correlation coefficients above 0.9 in experiment 1

TABLE II. Summary of correlation coefficients, r^2, and RMSEs (in dB) for each model and in experiments 1 and 2. The proposed model is the B-sEPSM. BE-sEPSM and BU-sEPSM are alternate outputs which consider only the better-ear, or only the binaural unmasking, respectively. The STEC model is from Wan *et al.* (2014) and the stBSIM is from Beutelmann *et al.* (2010). There are no performance values for experiment 3 because it consisted of only two data points.

Model	Exp. 1, r^2/RMSE	Exp. 2, r^2/RMSE
B-sEPSM	0.91/3.0	0.91/6.5
BE-sEPSM	0.83/4.4	0.91/4.1
BU-sEPSM	0.90/3.5	0.92/5.4
STEC	0.97/1.30	—
stBSIM	—	0.89/3.65

(Hawley *et al.*, 2004). However, the two models differ in a few important ways. Unlike the STEC, the B-sEPSM required a single parameter fit for the intelligibility transform for the whole experiment, rather than once per subcondition. In fact, the B-sEPSM, and sEPSM framework in general, requires a single parameter fit per speech material. In contrast, the STEC approach was validated using a different SII criterion (SII corresponding to 50% intelligibility) for each type and number of maskers. The generality of the sEPSM approach to model fitting was validated in the current study, as well as in Jørgensen and Dau (2011) and Jørgensen *et al.* (2013). Given the appropriate reference condition, which is typically in the presence of an SSN masker, the mr-sEPSM and its variants could account for a large range of processing or masker types, which means that the model requires less *a priori* knowledge about each condition. Another difference is that in the B-sEPSM, the BE and BU pathways are processed using similar time-frames, i.e., all pathways use the same multi-resolution approach to slice the time signals into segments. This means that the B-sEPSM can account for the monaural presentation of speech against a modulated masker because all pathways include short-term processing of the signals, and therefore the B-sEPSM would be compatible with the monaural mr-sEPSM. In contrast, only the BU pathway of the STEC considers a short-term process; the left- and right-ear pathways are applied to the long-term signals only. This is an important limitation of the STEC approach, considering the ability of the auditory system to extract information from BE glimpses, even if they shift across ears (Brungart and Iyer, 2012).

In experiment 2, the proposed model had a similar correlation coefficient as the stBSIM, but a slightly larger RMSE. Both the B-sEPSM and the stBSIM required a single parameter fit to convert the output of their decision metric to intelligibility. Unlike the B-sEPSM and the STEC, which explicitly separate the BE from the BU processes, the stBSIM implicitly includes the BE process in its closed-form calculation of the effective SNR [Beutelmann *et al.*, 2010, their Eq. (12)]. It would be possible, however, to create a BE-only version of the stBSIM by removing the ITD-related parameters from that equation, or conversely, to create a BU-only version of the model by removing the ILD-related parameters. However, this binding of the two binaural processes limits the feasibility of modifying the processes

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 70.113.67.182 On: Sun, 17 Jul 2016 23:24:33

independently, e.g., to use different time scales for the frame processing, or to introduce different amounts of sluggishness in each pathway (Culling and Summerfield, 1998; Culling and Mansell, 2013).

Neither the model of Lavandier and Culling (2010) nor any of its extensions was considered in the present study (Lavandier *et al.*, 2012; Collin and Lavandier, 2013). Of the extensions, only the one of Collin and Lavandier (2013) could possibly account for the masking release due to fluctuating maskers used in the majority of conditions considered in this study, because it is the only version that includes a short-term process. Those models are fundamentally limited because they cannot account for the effect of reverberation on the speech itself because they are not "signal-based," i.e., they do not use speech signals as targets, but rather rely on SSN as the target or on binaural room impulses. These simplifications make those model faster to compute than the proposed model as well as the STEC and the stBSIM, which makes them better tools for, e.g., acoustical room design but limits their applicability in certain scenarios.

Compared to the other models [STEC, stBSIM, Lavandier and Culling (2010), and even the binaural STI (Van Wijngaarden and Drullman, 2008)], the B-sEPSM avoids the need for the explicit frequency weighting from the SII. Instead, the frequency and modulation frequency weightings are limiting the processing to "audible" audio and modulation frequencies (Chabot-Leclerc *et al.*, 2014). Therefore, although the B-sEPSM includes the additional modulation-frequency dimension to the model framework, it reduces the number of fitted parameters required.

Overall, the modeling approach taken by the B-sEPSM, the STEC, and the BSIM did not differ largely. All three models combined a short-term EC process with time-frequency-specific cancellation parameters and a (short- or long-term) BE process. The main difference lay in the decision metric used by the B-sEPSM, namely, the SNR_{env} rather than the audio SNR, and the fact that the B-sEPSM included an envelope-domain audio-frequency-selective process.

B. Role of the decision metric

The SII-based models would fail in conditions with non-linear processing, such as noise reduction (Rhebergen *et al.*, 2009). The stBSIM as well as the model Collin and Lavandier (2013) are also fundamentally limited in that they cannot account for the effects of reverberation on the speech itself, because they do not use speech as target signal. Only the binaural STI model (Van Wijngaarden and Drullman, 2008), which uses the modulation power reduction after processing as the decision metric, could account for effects of modulation processing, but this approach is also limited because it cannot account for the intelligibility with modulated maskers. The B-sEPSM is the only binaural modeling framework that could account for multiple modulated maskers, reverberation on the target and maskers, as well as non-linear processing. Although these types of processing were not considered in the current study, the mr-sEPSM has been validated in such conditions (Jørgensen and Dau, 2011; Jørgensen *et al.*, 2013; Chabot-Leclerc *et al.*, 2014). No audibility-based model has been demonstrated to account for the change of intelligibility due to amplitude compression (Rhebergen *et al.*, 2009). Although the mr-sEPSM was not shown to account for the deleterious effect of amplitude compression on speech intelligibility, it could account for spectral subtraction, and, also to phase jitter, given the addition of an across-channel process (Chabot-Leclerc *et al.*, 2014).

C. Contributions of better-ear and binaural unmasking processes

The explicit separation of the BE and BU pathways in the B-sEPSM makes it possible to analyze their contributions separately. Moreover, the performance of those alternate models can be an indicator of the respective importance of the processes involved in binaural hearing. Overall, the BE- and BU-only simulations, denoted as BE-sEPSM and BU-sEPSM, respectively, showed good agreements between data and simulations. They are depicted as dotted and dashed lines, respectively, in Figs. 2–6. In experiment 1, the BE-sEPSM had an overall correlation coefficient of 0.83 and the BU-sEPSM a correlation coefficient of 0.90, which both compare favorably with the complete model's correlation of 0.91 (see Table II for overview). The performances were similar in experiment 2, with a correlation of 0.91 for the BE-sEPSM, 0.92 for the BU-sEPSM, and 0.91 for the complete B-sEPSM. In experiment 3, the BE-sEPSM model failed completely to account for the masking release due to ITDs, as expected, whereas the BU-sEPSM predicted the masking release. The performance of the BE-only model supported the idea that better-ear glimpsing, both in time and in frequency, can account for large parts of spatial release from masking (Brungart and Iyer, 2012; Culling and Mansell, 2013) in realistic conditions. Glyde *et al.* (2013) suggested this statement to be valid only if the maskers produced mostly energetic masking, i.e., did not cause any confusion between the target and the maskers. This is in contrast to conditions where informational masking may be dominant, such as with certain speech maskers. Therefore, the good performance of the BE-sEPSM can be attributed to the fact that the maskers considered in the present study may have provided a similar degree of informational masking (SSN, SMSSN, multi-talker babble, and reversed-speech).

The BU-sEPSM model performed equally well as the complete model (B-sEPSM) overall, and could account for the entire SRM in experiment 3. The difference in simulated SRT between the BE-sEPSM and the BU-sEPSM can be attributed to the fact that both models used the same "left-ear" reference for the fitting of the ideal observer. This discrepancy suggests that either they should be fitted separately, or that the processes should be modified as to produce the same SNR_{env} values in the same colocated condition. In experiments 1 and 2, the BU output "dominated" the B-sEPSM output, because its SNR_{env} values were larger than that of the BE-sEPSM (which leads to lower SRTs), as it is especially clear in Fig. 2. Also, the BU-sEPSM tended to predict a larger masking release than the BE-sEPSM (cf. Figs. 3 and 5). It is unclear if this dominance of the BU

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 70.113.67.182 On: Sun, 17 Jul 2016 23:24:33

pathway is an artifact of the modeling or if it is a property of the human binaural system. If the lower SRT predicted by the BU-sEPSM compared to the BE-sEPSM are modeling artifacts, then they could possibly be mitigated by the inclusion of sluggishness to the EC process [Culling and Summerfield (1998); Culling and Colburn (2000)] or by an increase of the EC jitters, which would limit its efficacy. Additionally, it may be that the constant short 20 ms windows of the EC process give the BU an advantage over the monaural pathways, where the multi-resolution approach is used. The EC window lengths could be adjusted or limited to restrict this advantage.

Some binaural models of speech intelligibility consider binaural unmasking as an additive process, while others do not. According to Culling and Mansell (2013), intelligibility benefits due to ILD and ITD seem to be additive. The modeling approach of Lavandier and Culling (2010) works under the same assumption that the total binaural advantage is the sum of the BE advantage and the advantage due to ITD processing (BMLD). In the model, only ILDs are considered in the BE pathway and only ITDs are considered in the BMLD pathway. The BSIM approach also indirectly uses this approach, where the ITD contributions can improve the SNR beyond the "better-ear" SNR (Beutelmann *et al.*, 2010). In contrast, the B-sEPSM and the STEC use a selection between the BU and BE, as if they are two separate processes and one of them can outperform the other in a given situation. In these two models, both ILDs and ITDs are considered in the BU pathway. Culling *et al.* (2004) studied the role of ILDs and ITDs using a subset of the conditions presented by Hawley *et al.* (2004). They considered the conditions with three speech or three SSN maskers, but presented binaural signals that had only ILDs, only ITDs, or were unmodified. They found the SRT patterns of the ITD-only and unmodified conditions to be similar, although the ITD-only condition had smaller differences between the spatial configurations. The ILD-only condition showed an SRM only when all maskers were on the right, otherwise the SRTs were the same as when all maskers were colocated with the target. For both masker types, considering the overall binaural advantage as the sum of the BE SRM and of the ITD SRM would lead to a large overestimation of the SRM in the unmodified condition. Therefore, in this condition, an "additive" binaural process is not appropriate and a selection process, such as in the B-sEPSM and STEC, seems more suitable.

D. Informational masking

The B-sEPSM predicted the correct SRM in experiment 1 with reversed-speech maskers [cf. Fig. 3 although simulated SRTs were lower than the data (cf. Fig. 2)]. A similar difference was observed with the SSEC and the STEC (Wan *et al.*, 2010, 2014) in the same condition. However, the models could not account for the increased thresholds observed when target and speech, or reversed-speech maskers, were colocated (Westermann and Buchholz, 2015b; Carlile and Corkhill, 2015). This limitation was even more clearly illustrated by Wan *et al.* (2014) in the conditions of Marrone *et al.* (2008), where the target was placed at 0° azimuth and speech or reversed-speech maskers were either colocated with the target or symmetrically placed around it. The models predicted SRTs lower than the data in the colocated condition because they could not account for the increased IM. In this case, IM is attributed to a failure in bottom-up grouping and streaming caused by target-masker similarities (Shinn-Cunningham, 2008). This is in contrast to the other portion of IM which can be attributed to top-down processes that cannot select the proper stream due to object similarity and target uncertainty (Shinn-Cunningham, 2008).

Being a purely bottom-up model, the B-sEPSM could only be sensitive to the similarity-based IM. However, the B-sEPSM has "perfect" segregation because of its access to the noisy mixture and to the maskers-alone signals and therefore cannot account for any IM. This means that the B-sEPSM requires fitting to a condition without IM, otherwise other simulated thresholds, where IM is not dominant, will be systematically elevated (e.g., in spatially separated conditions). On the converse, simulated SRTs in IM-dominated conditions will be too low if the B-sEPSM is fitted to an IM-free condition, which is the "default" approach for the mr-sEPSM framework. To account for the discrepancy between predicted and measured SRT in IM-dominated conditions, the B-sEPSM would require an estimate of the bottom-up confusion. Chabot-Leclerc *et al.* (2014) showed that it was possible to capture 7 of the 10 dB of SRM observed when a speech maskers was moved, on-axis, from 0.5 to 10 m away from the target in a reverberant environment (Westermann and Buchholz, 2015a) using the long-term sEPSM. Models based on the audio SNR (e.g., SII, BSIM) did not predict any SRM. Therefore, it seems that it is possible to capture some of the similarity/dissimilarity in the envelope-power representation which is not available in the audio domain. Consequently, it should be possible to evaluate the similarity between the speech and maskers using an estimate of the clean speech representation $[\hat{S} = (S + N) - N]$ and the maskers-alone representation in the envelope power domain. A simple "distance" or "contrast" estimate between the clean speech estimate and the maskers could be a promising measure of confusions. A more complex approach for estimating confusions would be to pair the B-sEPSM with a streaming model (e.g., Elhilali and Shamma, 2008; Christiansen *et al.*, 2014) and combine their outputs considering that there are more confusions in a one-stream percept than in a two-stream percept. It would be particularly interesting to apply this approach to the output of the binaural unmasking pathway considering that BE seems to be sufficient to account for SRM when there is no IM (Glyde *et al.*, 2013; Brungart and Iyer, 2012; Carlile and Corkhill, 2015).

VI. CONCLUSIONS

The B-sEPSM is a general model framework for predicting spatial release from masking in realistic and artificial conditions. It combines an explicit combination of better-ear and binaural unmasking processes using monaural implementations of the mr-sEPSM (Jørgensen *et al.*, 2013) and an EC process (Wan *et al.*, 2014). The B-sEPSM uses the

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 70.113.67.182 On: Sun, 17 Jul 2016 23:24:33

SNR_{env} as the decision metric and was shown to predict the SRT dependence on: the number of maskers, different masker types (SSN, SMSSN, babble, and reversed speech), the masker(s) azimuths, reverberation on the target and masker, and the ITD of the target and masker.

ACKNOWLEDGMENTS

The authors thank Rainer Beutelmann for providing the room impulse responses used in experiment 2 and Gustav Lőcsei for the material used in experiment 3. This research was supported in part by the National Science and Engineering Research Council of Canada (NSERC), Phonak, and the Technical University of Denmark.

ANSI (**1997**). ANSI S3.5, *American National Standard Methods for Calculation of the Speech Intelligibility Index* (American National Standards Institute, New York).

Auditec (**2006**). "CD101RW2, Audio CD," http://www.auditec.com (Last viewed 9/28/15).

Bernstein, J. G. W., and Grant, K. W. (**2009**). "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **125**, 3358–3372.

Bernstein, L. R., and Trahiotis, C. (**1996**). "The normalized correlation: Accounting for binaural detection across center frequency," J. Acoust. Soc. Am. **100**, 3774–3784.

Beutelmann, R., and Brand, T. (**2006**). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **120**, 331–342.

Beutelmann, R., Brand, T., and Kollmeier, B. (**2010**). "Revision, extension, and evaluation of a binaural speech intelligibility model," J. Acoust. Soc. Am. **127**, 2479–2497.

Blauert, J., Brueggen, M., Hartung, K., Bronkhorst, A. W., Drullmann, R., Reynaud, G., Pellieux, L., Krebber, W., and Sottek, R. (**1998**). "The AUDIS catalog of human HRTFs," J. Acoust. Soc. Am. **103**, 3082–3082.

Breebaart, J., van de Par, S., and Kohlrausch, A. (**2001**). "Binaural processing model based on contralateral inhibition. I. Model structure," J. Acoust. Soc. Am. **110**, 1074–1088.

Bronkhorst, A., and Plomp, R. (**1988**). "The effect of head-induced interaural time and level differences on speech intelligibility in noise," J. Acoust. Soc. Am. **83**, 1508–1516.

Brungart, D. S., and Iyer, N. (**2012**). "Better-ear glimpsing efficiency with symmetrically-placed interfering talkers," J. Acoust. Soc. Am. **132**, 2545–2556.

Carlile, S., and Corkhill, C. (**2015**). "Selective spatial attention modulates bottom-up informational masking of speech," Sci. Rep. **5**, 8662.

Chabot-Leclerc, A., Jørgensen, S., and Dau, T. (**2014**). "The role of auditory spectro-temporal modulation filtering and the decision metric for speech intelligibility prediction," J. Acoust. Soc. Am. **135**, 3502–3512.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.

Christensen, C. L. (**2005**). "Odeon room acoustics program, version 8.0," http://www.odeon.dk (Last viewed 5/28/15).

Christiansen, S. K., Jepsen, M. L., and Dau, T. (**2014**). "Effects of tonotopicity, adaptation, modulation tuning, and temporal coherence in 'primitive' auditory stream segregation," J. Acoust. Soc. Am. **135**, 323–333.

Collin, B., and Lavandier, M. (**2013**). "Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers," J. Acoust. Soc. Am. **134**, 1146–1159.

Culling, J. F., and Colburn, H. S. (**2000**). "Binaural sluggishness in the perception of tone sequences and speech in noise," J. Acoust. Soc. Am. **107**, 517–527.

Culling, J. F., Hawley, M. L., and Litovsky, R. Y. (**2004**). "The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources," J. Acoust. Soc. Am. **116**, 1057–1065.

Culling, J. F., Hawley, M. L., and Litovsky, R. Y. (**2005**). "Erratum: The role head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources [J. Acoust. Soc. Am. **116**, 1057 (2004)]," J. Acoust. Soc. Am. **118**, 552–552.

Culling, J. F., and Mansell, E. R. (**2013**). "Speech intelligibility among modulated and spatially distributed noise sources," J. Acoust. Soc. Am. **133**, 2254–2261.

Culling, J. F., and Summerfield, Q. (**1998**). "Measurements of the binaural temporal window using a detection task," J. Acoust. Soc. Am. **103**, 3540–3553.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997**). "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," J. Acoust. Soc. Am. **102**, 2892–2905.

Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (**2001**). "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment," Audiology **40**, 148–157.

Durlach, N. (**1963**). "Equalization and cancellation theory of binaural masking-level differences," J. Acoust. Soc. Am. **35**, 1206–1218.

Elhilali, M., and Shamma, S. A. (**2008**). "A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation," J. Acoust. Soc. Am. **124**, 3751–3771.

Ewert, S. D., and Dau, T. (**2000**). "Characterizing frequency selectivity for envelope fluctuations," J. Acoust. Soc. Am. **108**, 1181–1196.

Glyde, H., Buchholz, J., Dillon, H., Best, V., Hickson, L., and Cameron, S. (**2013**). "The effect of better-ear glimpsing on spatial release from masking," J. Acoust. Soc. Am. **134**, 2937–2945.

Hawley, M., Litovsky, R., and Culling, J. (**2004**). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," J. Acoust. Soc. Am. **115**, 833–843.

Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B. (**2010**). "Development and analysis of an international speech test signal (ISTS)," Int. J. Audiol. **49**, 891–903.

Houtgast, T., and Steeneken, H. J. (**1973**). "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acta Acust. Acust. **28**, 66–73.

IEC (**2003**). IEC60268-16, *Sound System Equipment—Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index* (International Electrotechnical Commission, Geneva, Switzerland).

ISO (**2005**). 389-7, *Reference Zero for the Calibration of Audiometric Equipment—Part 7: Reference Threshold of Hearing under Free-Field and Diffuse-Field Listening Conditions* (International Organization for Standardization, Geneva, Switzerland).

Jelfs, S., Culling, J. F., and Lavandier, M. (**2011**). "Revision and validation of a binaural model for speech intelligibility in noise," Hear. Res. **275**, 96–104.

Jørgensen, S., and Dau, T. (**2011**). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," J. Acoust. Soc. Am. **130**, 1475–1487.

Jørgensen, S., Ewert, S. D., and Dau, T. (**2013**). "A multi-resolution envelope-power based model for speech intelligibility," J. Acoust. Soc. Am. **134**, 436–446.

Lavandier, M., and Culling, J. F. (**2007**). "Speech segregation in rooms: Effects of reverberation on both target and interferer," J. Acoust. Soc. Am. **122**, 1713–1723.

Lavandier, M., and Culling, J. F. (**2010**). "Prediction of binaural speech intelligibility against noise in rooms," J. Acoust. Soc. Am. **127**, 387–399.

Lavandier, M., Jelfs, S., Culling, J. F., Watkins, A. J., Raimond, A. P., and Makin, S. J. (**2012**). "Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources," J. Acoust. Soc. Am. **131**, 218–231.

Levitt, H., and Rabiner, L. (**1967**). "Predicting binaural gain in intelligibility and release from masking for speech," J. Acoust. Soc. Am. **42**, 820–829.

Lőcsei, G., Hefting Pedersen, J., Laugesen, S., Santurette, S., Dau, T., and MacDonald, E. N. (**2015**). "Lateralized speech perception, temporal processing and cognitive function in NH and HI listeners," presented at the *Speech in Noise Workshop*, Copenhagen, Denmark.

Loizou, P. C. (**2007**). *Speech Enhancement: Theory and Practice*, 1st ed. (CRC, Boca Raton, FL).

Marrone, N., Mason, C. R., and Kidd, G. (**2008**). "Tuning in the spatial dimension: Evidence from a masked speech identification task," J. Acoust. Soc. Am. **124**, 1146–1158.

Nielsen, J. B., Dau, T., and Neher, T. (**2014**). "A Danish open-set speech corpus for competing-speech studies," J. Acoust. Soc. Am. **135**, 407–420.

Plomp, R. (**1976**). "Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise)," Acustica **34**, 200–211.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 70.113.67.182 On: Sun, 17 Jul 2016 23:24:33

Rennies, J., Brand, T., and Kollmeier, B. (**2011**). "Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet," J. Acoust. Soc. Am. **130**, 2999–3012.
Rhebergen, K. S., and Versfeld, N. J. (**2005**). "A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," J. Acoust. Soc. Am. **117**, 2181–2192.
Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2009**). "The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise," J. Acoust. Soc. Am. **126**, 3236–3245.
Rothauser, E., Chapman, W., Guttman, N., Nordby, K., Silbiger, H., Urbanek, G., and Weinstock, M. (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.
Shinn-Cunningham, B. G. (**2008**). "Object-based auditory and visual attention," Trends Cogn. Sci. **12**, 182–186.
Van Wijngaarden, S., and Drullman, R. (**2008**). "Binaural intelligibility prediction based on the speech transmission index," J. Acoust. Soc. Am. **123**, 4514–4523.
Verhey, J. L., Dau, T., and Kollmeier, B. (**1999**). "Within-channel cues in comodulation masking release (CMR): Experiments and model predictions using a modulation-filterbank model," J. Acoust. Soc. Am. **106**, 2733–2745.
Wagener, K., Kühnel, V., and Kollmeier, B. (**1999**). "Development and evaluation of a German sentence test I: Design of the Oldenburg sentence test," Z. Audiol. Audiol. Acoust. **38**, 4–15.
Wagener, K. C., and Brand, T. (**2005**). "Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters," Int. J. Audiol. **44**, 144–156.
Wan, R., Durlach, N. I., and Colburn, H. S. (**2010**). "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers," J. Acoust. Soc. Am. **128**, 3678–3690.
Wan, R., Durlach, N. I., and Colburn, H. S. (**2014**). "Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers," J. Acoust. Soc. Am. **136**, 768–776.
Westermann, A., and Buchholz, J. M. (**2015a**). "The effect of spatial separation in distance on the intelligibility of speech in rooms," J. Acoust. Soc. Am. **137**, 757–767.
Westermann, A., and Buchholz, J. M. (**2015b**). "The influence of informational masking in reverberant, multi-talker environmentsa)," J. Acoust. Soc. Am. **138**, 584–593.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 70.113.67.182 On: Sun, 17 Jul 2016 23:24:33